Comparative Evaluation of Touch-Based Input Techniques for Experience Sampling on Smartwatches

Document license:
CC BY

DOI:
[10.1145/3626705.3627768](10.1145/3626705.3627768)

Document status and date:
Published: 03/12/2023

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Comparative Evaluation of Touch-Based Input Techniques for Experience Sampling on Smartwatches

Panos Markopoulos
p.markopoulos@tue.nl
Industrial Design Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

Alireza Khanshan
a.khanshan@tue.nl
Industrial Design Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

Sven Bormans
s.r.bormans@student.tue.nl
Industrial Design Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

Gabriella Tisza
g.tisza@tue.nl
Industrial Design Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

Ling Kang
l.kang@student.tue.nl
Industrial Design Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

Pieter Van Gorp
p.m.e.v.gorp@tue.nl
Industrial Engineering and
Innovation Sciences Department of
Eindhoven University of Technology
Eindhoven, The Netherlands

ABSTRACT

Smartwatches are emerging as an increasingly popular platform for longitudinal in situ data collection with methods often referred to as experience sampling and ecological momentary assessment. Their small size challenges designers of relevant applications to ensure usability and a positive user experience. This paper investigates the usability of different input techniques for responding to in situ surveys administered on smartwatches. In this paper, we classify different input techniques that can support this task. Then, we report on two user studies that compared different input techniques and their suitability at two levels of user activity: while sitting and while walking. A pilot study ($N = 18$) examined numeric input with three input techniques that utilize common features of smartwatches with a touchscreen: Multi-Step Tapping, Bezel Rotation, and Swiping. The main study ($N = 80$) examined numeric input and list selection including in the comparison two more techniques: Long-List Tapping and Virtual Buttons to scroll through options. Overall, we found that whether users are seated or walking did not affect the speed or accuracy of input. Bezel rotation was the slowest input technique but also the most accurate. Swiping resulted in most errors. Long-List Tapping yielded the shortest reaction times. Future research should examine different form factors for the smartwatch and diverse usage contexts.

CCS CONCEPTS

• **Human-centered computing** → **Interaction devices**; **Empirical studies in interaction design**; *Ubiquitous and mobile devices.*

KEYWORDS

Interaction design, Smartwatch application, Context-aware experience sampling

1 INTRODUCTION

The increasing adoption of smartwatches means that a growing number of people wear them and can use them for performing tasks across contexts and at different times of the day. This development makes them a particularly suitable platform for surveying user opinions and attitudes in situ, as they go about their real-life activities in a variety of contexts. We are particularly interested in supporting wESM [14] approaches which use wearables as notification and input devices for the Experience Sampling Method (ESM), for surveying people's thoughts, feelings, and behaviors repeatedly, at various times of the day and over sustained study periods [9, 22]. While ESM originates from the field of psychology, it is a method that has been widely adopted in various scientific areas and application domains, and among those in human-computer interaction, as a way to study user needs and user experiences 'in the wild' [5, 8, 42].

ESM and related methods for in situ data collection, also described as ecological momentary assessment [38], or just-in-time ecological momentary assessment [36], aim to ensure reliable self-report and high response rates by reducing the need for retrospection. However, data collection is hampered by the need for the participant to carry a device, to use it in different perhaps unsuitable contexts, and by expecting respondents to enter data while interrupting other activities they happen to be engaged in. Wearables and more specifically smartwatches can alleviate this problem as they can be ready at hand and potentially worn continuously during a study period [6]. Furthermore, they allow the collection of objective sensor data using embedded sensors in the device.

Here we are interested specifically in self-report data and the efficacy of different user interfaces for this purpose. Smartwatches have a minimal interface for data entry [24]. Current smartwatches

typically rely on touchscreen-based interaction confined to a tiny screen, for which it is challenging to ensure usability. Research in this area has explored user interactions with such devices looking at patterns of behavior [41], opening up the design space for smartwatch interfaces, e.g., for displaying notifications and information [44], or text input through virtual keyboards which is particularly challenging on small screens [18, 23, 26, 33, 40]. Other research has examined menu structures and the organization of information so that it can be easily located and selected on a smartwatch interface, e.g., [27, 31].

Previous research has investigated wESM implementation [11, 15, 24, 42, 45]. However, the design space of input techniques suitable for surveying user attitudes and opinions remains to be comprehensively assessed. Our research aims to fill this gap and so inform the user interface design of wESM applications. To characterize the relevant design space we first classified available input techniques. Then we compared user performance for different input techniques during two studies. In a pilot study (N=18), we compared three different input techniques (multi-step (MS) tapping, swiping, and bezel rotation) for entering numbers when the user is seated and when the user is walking. The pilot study exposed some usability limitations of our implementation of these techniques which were subsequently corrected. In the main study (N=80)we extended the collection of input techniques assessed with two different ways for list selection, namely Long-List Tapping (or LL-tapping) and Virtual Button (or V-button). We then assessed how our five techniques could support numeric input (with relatively small numbers) and list selection (again from a relatively short collection of options). The first study suggested MS-tapping to be superior. The second and more extensive study found that the Bezel rotation was the slowest input technique but also the most accurate while swiping resulted in the most errors. Long-List Tapping was the fastest if both tasks are considered. These results contribute a clear guideline for the instrumentation of wESM studies to prefer LL-tapping for speed and Bezel rotation for accuracy.

The following section summarizes earlier research on input techniques for smartwatches. We then go on to describe the two studies detailing methods and results, and conclude by discussing our results and their implications for future research.

2 RELATED WORK

Technologies to support ESM studies have closely tracked the developments in mobile technologies from the days of personal digital assistants, e.g., see [3], to the current almost total adoption of smartphones, e.g., [29, 42]. Recently researchers have turned their attention to smartwatches, for the reasons mentioned above regarding their increased availability and unobtrusiveness compared to smartphones. Intille et al. implemented μEMA as an extension to smartphones that delivered prompts on the smartwatch as well as concise versions of ESM questions [12]. Compared to running an ESM protocol on a smartphone, they observed significantly higher compliance, completion, and first prompt response rates when running it on a smartwatch, and participants found it less distracting. Technology development efforts have resulted in different smartwatch-based frameworks for real-time, online assessment

and activity monitoring [14, 16]. Hafiz et al. [10] found a strong correlation between the data gathered with cognitive assessment tests administered via their purpose-made smartwatch application and computer-based tests administered in a lab, which indicates that reporting can be reliable even for tasks that are quite demanding.

Smartwatches support a variety of input techniques via touch, voice, or gesture [4, 46]. The small touchscreen of smartwatches makes input tasks challenging, for which interaction techniques need to be adapted or specially developed, e.g., [30, 31]. Touch input tasks are challenging and users are most efficient when fixing the position of their dominant hand in relation to the smartwatch screen by using their thumb or their thumb and middle finger [41]. Their ability to interact with smartwatches is affected by their activity (e.g., sitting, walking, or running) and encumbrance (whether their hands are occupied when they need to interact with the smartwatches) [39], and this may influence the response rate in experience sampling studies [15]. An investigation of smartwatch touch-only interaction for the tasks of target selection, panning, zooming, and flicking concluded that two-finger input should be avoided, and that flick- and tap-based interactions are more robust for different levels of physical activity and encumbrance [39]. A usability evaluation of different smartwatch menu designs in [27] found that menu selection is more efficient in a list rather than a grid layout, and a hierarchical organization of the menu is superior both in efficiency, and overall satisfaction. Text input is particularly challenging but can be facilitated by progressively zooming into different portions of the keyboard and by offsetting the display of the keyboard in relation to the user's finger, which would otherwise cover the soft keyboard during text entry [23].

Specifically for conducting in situ surveys, user interaction consists of relatively well-defined tasks that involve presenting a prompt or question to the user and collecting their self-report. A comprehensive classification of common input tasks for ESM studies presented in [42] includes tasks such as rating Likert scales, selecting options on radio buttons or checkboxes, text input, and sliders. Here we focus on two of these that together cover most needs of the tasks that users need to perform during ESM studies: a) entering a small number (typically one digit) can support rating scales that are typically used to report on an attitude or experience of the user and b) selecting from a list of options which can be used for scoring on nominal variables, e.g. identifying emotions or situations by their name. For rating scales there is as yet no specific design guideline for smartwatches [2, 12, 43, 45]. For list selection some guidance can be derived from studies on organizing menus on smartwatches, e.g., [27, 31] though in the case of ESM, we are interested in shorter menus rather than the generic case of browsing larger collection of information items. We investigate the user performance in terms of accuracy and speed, with five different input techniques at two different activity modes (walking and sitting). We also explore whether the suitability of different techniques varies with context, which would suggest the need for a context-adaptive user interface. Earlier research has demonstrated how context awareness can help optimize user interfaces in mobile and wearable devices, e.g., see [19]. Such an adaptation of wESM user interfaces to context may be an interesting approach to ensure increased adherence to ESM protocols, as the different contexts and

physical activity are known to influence responses and adherence to ESM protocols, e.g., see [15, 25].

3 CLASSIFICATION OF INPUT TECHNIQUES FOR RATING SCALES AND LIST SELECTION

Two input tasks are particularly relevant for supporting ESM studies where users report on moods, feelings, e.g., [22], contexts, and activities, e.g., [5, 13]. One task is rating numeric scales, which amounts to entering a digit within that range, e.g., one to seven. Another is list selection, where a list of options is offered to the user to choose from. As users are asked to respond frequently to ESM prompts, researchers avoid offering too many options. So where earlier studies examining how smartwatch interfaces can support users to locate and choose from a large number of options, e.g., 15 to 240 items in [30], or 40 items as in [28, 31], a typical list selection task for self-report in ESM studies may be lower than 12, e.g., see [5, 13], where the options offered were kept low to limit the effort needed for frequent self-report. Reviewing related literature we identify the following input techniques that can support the two selected tasks on smartwatches illustrated in Fig. 1:

(1) *Multi-step (MS) tapping*: The user answers each question in multiple steps by first choosing a range where their answer is in and continuing so until the desired answer is selected [45].

(2) *Single-step (SS) tapping*: The user sees all the options displayed on a single screen and taps to select one [12, 20, 32, 45].

(3) *Long-list (LL) tapping*: A list menu is displayed on the smartwatch screen, containing all questionnaire options. Users can swipe to scroll the list and tap to select the desired option.

(4) *Arm movement*: the user moves their entire arm to scroll through the menu or to choose a particular answer [45].

(5) *Wrist rotation*: The user rotates their wrist to scroll through the menu or to choose a particular answer [45].

(6) *Swiping*: The user continuously touches the screen toward one direction [45].

(7) *Virtual (V) buttons*: Users can tap buttons to paging through options one at a time. These virtual buttons are visually presented on the screen and respond to the user's touch input for option selection.

(8) *Voice input*: The user speaks out the desired number or option to the microphone of the device [1, 17].

(9) *Flicking*: The user rotates their wrist abruptly clockwise or counter-clockwise [45].

(10) *Button press*: The user presses the buttons on the frame of the watch to navigate between options and taps to select the currently displayed option [17, 37].

(11) *Bezel rotation*: The user rotates the ring around the watch to shift through options, and taps in the central region of the dial to select the value it displays at that moment [21]. To select among a large number of items, this can be combined with the list selection, using the bezel to select list segments and list selection to locate an individual item as in [31].

(12) *Sliding*: Sliding is a variant of swiping where the user swipes in a circular motion (clockwise or counter-clockwise) on the screen to select an answer [45]

(13) *Drawing*: The user slides their finger on the screen and draws the desired input such as numbers or simplified drawings [45]. This amounts to handwriting with the finger on the smartwatch screen.

4 PILOT STUDY: NUMERIC INPUT WITH MULTI-STEP TAPPING, BEZEL ROTATION, AND SWIPING

Our pilot aimed to explore how different input techniques affect user performance when reporting on a rating scale, as is common in ESM protocols. Specifically, we compared three of the input techniques described above, namely Bezel Rotation, MS-Tapping, and Swiping, for answering survey questions implemented on the same hardware and operating system. These three techniques were chosen based on earlier research. MS tapping has been found to be more efficient than single-step tapping [45] and it can accommodate multiple options thanks to its hierarchical organization. An earlier study [31] found that circular selection as in the Bezel rotation outperforms traditional smartwatch list interfaces in terms of user preference and task completion time. On the other hand, we considered swiping to be particularly relevant when ESM respondents are moving, as it requires less precise gestures. We designed and developed an ESM smartwatch application supporting these three input techniques and the procedures and measurements of the experiment. We measured the time taken to answer questions and whether the responses were correct. In order to assess how the level of physical activity may alter the suitability of different input methods, we compared user performance in these interactions while walking and sitting. In order to examine whether user interfaces for wESM should be adaptive, we were particularly interested to note if different interfaces are suitable for different levels of activity of users.

4.1 Methods

We conducted a within-subject, 3 (input techniques) × 2 (activity level) experiment with a crossover design, with $N = 18$ participants recruited through convenience sampling. The study was approved by the university ethics board. Participants did not receive any compensation for their participation in the study, and informed consent was obtained before the study took place.

4.1.1 Independent variables. We had two variables, the input technique (Bezel Rotation, MS-Tapping, and Swiping) and the activity level (sitting or walking).

4.1.2 Dependent variables. We measured the time taken to respond to the survey questions as logged by the software as a measure of the efficiency of user input and the number of incorrect answers as a negative measure of accuracy. We divided the number of incorrect responses by the number of all responses. Accordingly, higher average values indicate a higher error rate, and thus, a lower accuracy.

4.1.3 Hypotheses. Based on the literature review we expected that:

(1) The reaction time and the error rates will differ between the input techniques.

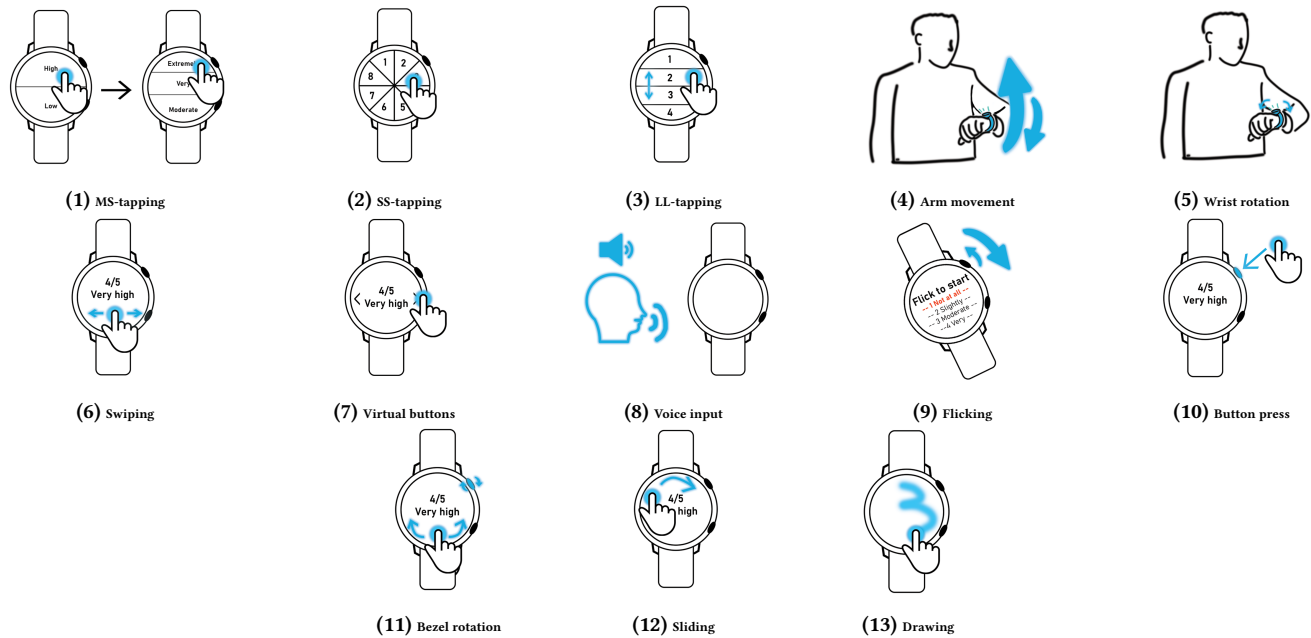(2) The reaction times and the error rates will differ when walking or sitting.

(1) MS-tapping   **(2)** SS-tapping   **(3)** LL-tapping   **(4)** Arm movement   **(5)** Wrist rotation

(6) Swiping   **(7)** Virtual buttons   **(8)** Voice input   **(9)** Flicking   **(10)** Button press

(11) Bezel rotation   **(12)** Sliding   **(13)** Drawing

Figure 1: Overview of input techniques on (circular) smartwatches

(3) There will be an interaction effect between the activity level and the input techniques, meaning that observed differences in reaction times and error rates for the different input techniques will be influenced by whether the users are walking or sitting.

4.1.4 Materials and Measures. The experimental protocol was supported by a purpose-made application developed for the Samsung Galaxy Watch Active 2 [35] (Tizen OS 5.5) using the Tizen Web API and the Tizen Advanced UI (TAU) framework [34], written in JavaScript. The source code of the app is available at https://github.com/khnshn/interaction-styles-evaluation. The application asks the user to respond to survey questions using the three input techniques compared. The questions, which are similar in structure to those asked in ESM surveys, are offered in a randomized order to limit the learning effect and expectancy bias. The application records the answers and the time users take to answer. The interface was designed to be minimal to avoid distracting users with unrelated design features and to avoid potentially biasing the experimental results. To ensure usability, all three input techniques implemented the design guidelines reported in [39], and [27]. Figure 2 provides an overview of the application and instances from the three input techniques evaluated.

4.1.5 Participants. Eighteen participants were recruited with convenience sampling, with ages ranging between 19 and 61 (*mean =* 34.11, *SD* = 17, 61) years. No participant suffered from poor vision.

4.1.6 Procedure. Each participant interacted with the software and was asked to enter data once while sitting in front of a desk and once again while walking in a large open office environment with relatively low occupancy and enough area for participants to walk unencumbered. To avoid learning effects we adopted a

crossover design, in which we asked half of the participants to perform the experiment at first while walking and then while sitting, and the other half vice versa. Moreover, a sequence of 18 questions on the smartwatch and their corresponding input technique was randomized for each experiment to avoid expectancy-related bias. In order to avoid influencing performance by how each participant would find a question easy or difficult to answer, we asked short simple questions with trivial answers. During the experiment, two types of questions were asked. The first type comprised 15 questions with numerical answers (Appendix A). Each question presented ten options, ranging from one to ten. The second type consisted of five questions with textual answers (Appendix A), where each question included ten different country names as options. For example, "Can you insert number 5?" or One+One=?" to minimize the cognitive load for deciding on what to respond and minimize the influence of cognitive processing on the performance of the input task. We also made sure the expected answers would cover the range of possible responses, to avoid a contrived comparison (e.g., when the expected answer would always be to the lower end of the range). Then, we measured the task completion time and errors per input technique. After interacting with the smartwatch, we conducted structured interviews with the participants to understand their subjective preferences. The interview protocol is shown in Appendix A.

4.1.7 Data Analysis. For the data analysis, IBM SPSS Statistics software version 27 was used. To test the hypotheses, a number of Repeated Measures of two-way ANOVA was conducted.

4.2 Results

4.2.1 Effects of activity level and input technique on reaction times. The mean reaction times measured in milliseconds for the two activity levels (i.e., sitting and walking) along the three input techniques

(1) Starting screen

(2) Question

(3) MS-tapping (step 1: all options)

(4) MS-tapping (step 2: e.g., options after selecting [5-7])

(5) Bezel rotation

(6) Swiping

(7) Completion (after a question)

(8) Completion (after the questionnaire)

Figure 2: Screenshots of the smartwatch application guiding participants through the study procedure and implementing the three input techniques we compare

are shown in Table 1. Fig. 3 displays the average reaction time along the three input techniques. The separate lines indicate the activity levels (i.e., sitting and walking conditions).

	Sitting	Walking
Bezel Rotation	5458 (SD = 3128)	5151 (SD = 2987)
MS-Tapping	2910 (SD = 1997)	2726 (SD = 1008)
Swiping	3625 (SD = 1844)	4011 (SD = 1879)

Table 1: Mean reaction time values (milliseconds) for the input techniques for the sitting and walking conditions.

We conducted repeated measures of two-way ANOVA to assess the effect of activity level on the reaction time for the three input techniques. The main effect (i.e., within-subject effect) of the activity level was not significant ($p = 0.786$, $F(1, 17) = 0.076$, Partial $\eta 2$ = 0.004), suggesting that, overall, participants were just as fast whether seated or walking. The main effect of the input technique was significant ($p < 0.001$, $F(2, 34) = 38.705$, Partial $\eta 2 = 0.695$). A posthoc analysis with a Bonferroni correction determined that there was a significant difference between all input technique pairs (pBezel R − MS-Tapping < 0.001; pBezel R − Swiping < 0.001; pMS-Tapping-Swiping = 0.008). MS-Tapping was the fastest followed by Swiping and Bezel Rotation. Furthermore, the repeated measures
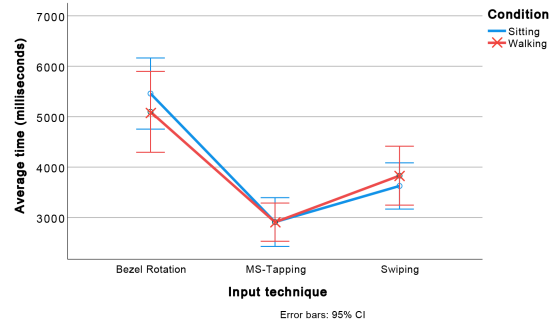

Figure 3: Reaction times while sitting and walking for the three input techniques.

ANOVA indicated a non-significant interaction effect ($p = 0.470$, $F(2, 34) = 0.773$, Partial $\eta 2= 0.043$). Thus, the differences in efficiency noted between the input techniques are similar when walking or seated.

4.2.2 Effects of activity level and input technique on error rates. The average values along the activity levels and three input techniques are shown in Table 2.

	Sitting	Walking
Bezel Rotation	0.28	0.20
MS-Tapping	0.06	0.19
Swiping	0.19	0.15

Table 2: Average error rates for the three input techniques for the sitting and walking conditions.

MS-Tapping resulted in the lowest error rate in the sitting condition, while Swiping resulted in the lowest error rate in the walking condition. On the other hand, most errors in both conditions were made with the Bezel Rotation input technique. Fig. 4 displays the average error rates along the three input techniques per activity level.
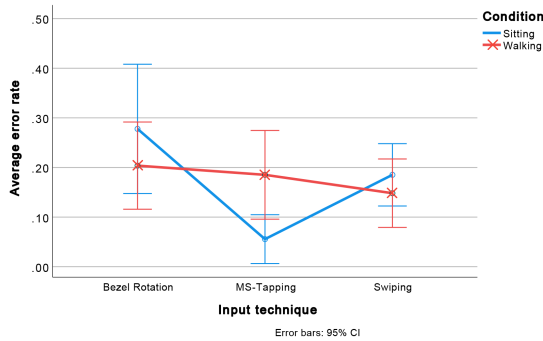

Figure 4: Average error rates in the sitting and walking conditions for the three input techniques.

A repeated measures two-way ANOVA did not find a significant main effect of the activity level (p = 0.163, F(1,17) = 2.125, $\eta 2$ = 0.111), suggesting that overall participants were just as accurate whether seated or walking. However, there was a significant main effect of the input technique on the error rate (p = 0.004, F(2, 34) = 6.384, Partial $\eta 2$ = 0.273). A post-hoc analysis with a Bonferroni correction determined that there was a significant difference (p = 0.026) between Bezel Rotation and MS-Tapping, the latter having a significantly lower error rate. Other pairwise comparisons were non-significant. Furthermore, the two-way ANOVA indicated a non-significant interaction effect (p = 0.059, F(2, 34) = 3.085), Partial $\eta 2$ = 0.154). In other words, whether the participant was sitting or walking did not influence the differences in the error rates found for the input techniques.

4.3 Qualitative data

Most participants commented that they favored MS-Tapping for being the quickest input technique while resulting in the least burden and fewest errors. They found that it gave them a clearer overview of the possible answers and felt it required less precise

action compared to the other techniques. However, they also found it the least visually appealing. The Bezel Rotation received mixed feedback. It was found convenient since it gave a detailed overview of all the possible answers while facilitating a back-and-forth transition between the choices. However, for users inexperienced with smartwatches, the interface did not clearly indicate what type of interaction was required to navigate through the options and how to select an answer. The swiping interface was considered to be sufficient and functional but was favored less than MS-Tapping, while it was found more convenient than the Bezel Rotation. Notably, 17 out of 18 participants mentioned that they experienced it as more effortful while walking, which was not reflected in the objective measures of performance or errors above. Additionally, the active area for tapping to select a response was found to be too close to the bezel, so some participants mentioned how they accidentally selected a response when they intended to rotate the bezel. A potential limitation we noted in this pilot study, is that there was a very high spread of ages in our convenience sample, which included young students as well as people close to retirement and few people in the range between. We compared the performance and preferences of these two subgroups and found that they diverged, potentially confounding our results. Another limitation concerns the implementation of the swiping interface. We observed that participants often mistook the arrows in the swiping screen in Fig. 2.6, which were purposed to solely provide a visual guide to indicate the swiping direction. Hence, they often tapped on those (non-interactive) icons hoping to scroll to the next value left or right rather than swiping, which led to longer interaction times. We considered this to be a limitation of our implementation of the input style rather than an inherent limitation of swiping as an input technique. For this, we decided to improve the application and repeat the comparison.

5 MAIN STUDY: NUMERIC INPUT AND LIST SELECTION WITH FIVE INPUT TECHNIQUES

To address the limitations of the pilot study we improved the application. First, we removed the arrows from the swiping interface to avoid them being mistaken for on-screen buttons. Considering how users intuitively interpreted these buttons we also decided to implement the V-Button input method, where the virtual buttons indicating the direction of scrolling could be pressed to page through the list of choices. We also increased the separation of the bezel rotation dial from the area for tapping to confirm a response. Moreover, to cover more of the input tasks needed for wESM, we extended our study to include list selection, i.e. to choose from a list of strings representing names or phrases. This is useful in ESM studies to indicate responses that require providing categorical and ordinal responses, e.g., to indicate moods, preferences, etc. Following the results of [27] we should expect a list of items to choose from to be a more efficient user interface than other organizations such as a two-dimensional grid. We thus extended our smartwatch app with the Virtual Button and the LL-tapping input techniques to test and evaluate whether reaction times and error rates are influenced by the activity level (see Fig. 5). We also included in the application list selection tasks that used the three initial input techniques. Finally,

to eliminate the spread in ages, we restricted the age of participants to only include students.

We conducted a within-subject, 5 (input techniques) × 2 (activity level) experiment with a crossover design, with $N = 80$ participants recruited through convenience sampling. The procedures and the measures in this study were similar to those in the pilot study described in section 4.1.

5.1 Methods

5.1.1 Participants. In this study, 80 students (43 male, 37 female) participated between ages 18 to 26 ($mean = 25.51, SD = 2.158$). The sample size was determined a priori based on a target of 0.8 power, with at least medium effect sizes expected (f=0,4), using statistical power analysis on the G*Power application [7]. We narrowed our sampling frame to students in our university to avoid potential confounding effects relating to age and fluency with mobile, touch-based interactions. Participants did not have vision problems. Prior to the study, 24 participants reported having previous experience with using smart watches, from which 6 participants had used a Samsung smartwatch before; the rest of the participants were novices in this respect.

5.1.2 Procedure. The experiment was conducted on the campus of the [university], in the library building. Participants completed both conditions (i.e., sitting and walking) in a single setting. In the sitting condition participants were asked to take a seat to answer the questions, while in the walking condition, they were asked to walk around while responding to the questions. The participants underwent the experiment individually, and their questionnaire responses and time data were automatically recorded on the smartwatch. The experiment lasted approximately 12 minutes and consisted of the following steps: i) introduction to the study and obtaining consent, ii) introduction to the interaction methods, iii) experiment, iv) interview. During the experiment, two types of questions were asked. The first type comprised of 15 objective questions with numerical answers (Appendix A). Each question presented ten options, ranging from 1 to 10. The second type consisted of 5 objective questions with textual answers (Appendix A), where each question included ten different country names as options. In the experimental setting, all questions were presented in a randomized sequence. After the experimental tasks were completed we conducted structured interviews with all participants (n = 80) to gather their perspectives on the five interaction methods (i.e., Bezel rotation, MS-tapping, LL-tapping, Swipe, and V-Button. The interview comprised of the following six questions: i) How old are you?, ii) Do you have earlier experience with using a smartwatch?, iii) Which interaction style do you like the most? Why?, iv) Which interaction style do you hate the most? Why?, v) What is your opinion on each interaction style?, and vi) What is your opinion about each interaction style in reflection of the sitting and walking condition?

5.1.3 Independent variables. We had two variables, the input technique (Bezel Rotation, MS-Tapping, Swiping. LL-tapping, V-Button) and the activity level (sitting or walking).

5.1.4 Dependent Variables and Hypotheses. The dependent variables were the same as with the pilot study, our measures were the

	Sitting	Walking
Bezel Rotation	5691 (SD = 2052)	5269 (SD = 1603)
LL-Tapping	2816 (SD = 1217)	2866 (SD = 2026)
MS-Tapping	3119 (SD = 1033)	3189 (SD = 1462)
Swiping	3721 (SD = 1516)	3738 (SD = 1138)
V-Button	3036 (SD = 1264)	3123 (SD = 995)

Table 3: Mean reaction time values (milliseconds) for the input techniques for the sitting and walking conditions with numeric input.

time taken to respond to survey questions and the number of errors made.

5.1.5 Hypotheses. The hypotheses were the same as for the pilot study, but this time for the numeric and the list selection tasks, and using all five input techniques.

5.2 Results

In order to address the research questions and investigate the expected interaction effect between the sitting and walking conditions and the input techniques, we conducted a number of 2-way repeated measures ANOVA tests.

5.2.1 Effects of activity level and input technique on reaction time for numeric input. The mean reaction time values (milliseconds) for the two conditions (i.e., sitting and walking) with numeric input, compared between the five input techniques are shown in Table 3. Based on the average values we note that the LL-Tapping had on average the lowest reaction time for numeric input, while the Bezel Rotation required the highest average reaction time in both the sitting and the walking condition. Fig. 6 displays the average reaction times along the five input techniques. The separate lines indicate the two activity levels(i.e., sitting and walking).

A repeated measures ANOVA was conducted to assess the effect of the activity level on the reaction time for the five input techniques for numeric input. The main effect of the activity level was not significant (p = 0.785, $F_{(1,79)} = 0.075$, $\eta2 = 0.001$), suggesting that overall participants were just as fast whether walking or seated. The main effect of the input technique was significant (p < 0.001, $F(4, 316) = 99.053$, Partial $\eta 2 = 0.556$). A post-hoc analysis with a Bonferroni correction indicated that there was a significant difference between Bezel Rotation and all other input techniques (all p < 0.001) and that there was a significant difference between Swiping and other input techniques (all p ≤ 0.001), both of them resulting in longer reaction times than the other three input techniques. Other pairwise comparisons were non-significant. The interaction effect was not significant (p = 0.381, $F(4, 316) = 1.051$), Partial $\eta 2 = 0.013$), too. In other words, the differences in efficiency between the input techniques were independent of the activity level.

5.2.2 Effects of activity level and input technique on reaction times for list selection. The mean reaction time values (milliseconds) for the two conditions (i.e., sitting and walking) while using each of the five input techniques for list selection tasks are shown in Table 4.

(1) Virtual buttons

(2) Swiping (no indicator buttons)

(3) LL-tapping

Figure 5: Screenshots of the additional input techniques developed for the main study
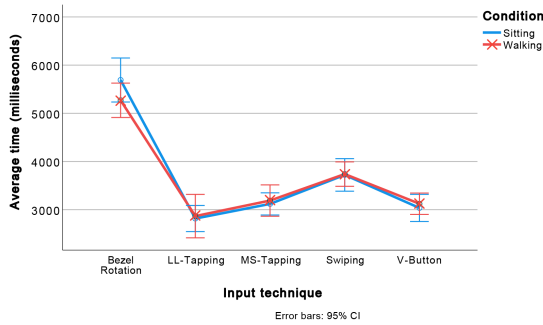

Figure 6: Reaction times for the sitting and walking conditions for the different input techniques for numeric input.
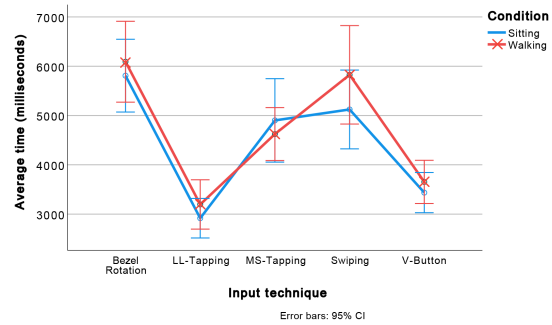

Figure 7: Reaction times in the sitting and walking conditions for list-selection tasks

	Sitting	Walking
Bezel Rotation	5808 (SD = 3317)	6091 (SD = 3686)
LL-Tapping	2916 (SD = 1802)	3195 (SD = 2248)
MS-Tapping	4901 (SD = 3804)	4623 (SD = 2412)
Swiping	5122 (SD = 3591)	5825 (SD = 4488)
V-Button	3435 (SD = 1832)	3652 (SD = 1968)

Table 4: Mean reaction time values (milliseconds) for list selection for the sitting and walking conditions.

For list selection tasks LL-Tapping had on average the lowest reaction time, while the Bezel Rotation required the highest average reaction time in both the sitting and the walking conditions. Fig. 7 displays the average reaction time along the five input techniques. The separate lines indicate the activity level (i.e., sitting and walking).

A repeated measures ANOVA did not reveal a significant main effect of the activity level ($p = 0.289$, $F(1, 79) = 1.141$, Partial $\eta 2 = 0.014$), suggesting that overall participants were just as fast whether seated or walking. As above, the main effect of the input techniques

was significant ($p<0.001$, $F(4, 316) = 27.378$, Partial $\eta 2 = 0.257$). Post-hoc pairwise comparisons for all pairs ($p < 0.05$) except between the Bezel Rotation and Swiping ($p = 1.000$), LL-Tapping and V-Button ($p = 0.483$), and MS-Tapping and Swiping ($p = 0.619$). Furthermore, the variance analysis indicated a non-significant interaction effect ($p = 0.724$, $F(4, 316) = 0.516$, Partial $\eta 2 = 0.006$), too. In other words, the differences in efficiency between the input techniques were independent of the activity level.

5.2.3 Effects of activity level and input technique on error rates for numeric input. The average error rates are calculated as the number of false responses over the total number of responses, and it shown in Table 5. We observe that the Bezel Rotation had on average the lowest error rate while Swiping had the highest, in both the sitting and the walking conditions. Fig. 8 displays the average error rates for the five input techniques. The separate lines indicate the level of activity (i.e., sitting and walking).

A repeated measures ANOVA did not find a significant main effect of the level of activity on the error rates for numeric input ($p = 0.070$, $F (1, 79) = 3.380$, Partial $\eta 2 = 0.041$), suggesting that, overall, participants were just as accurate whether walking or seated. The main effect of the input technique was significant ($p = 0.009$, $F(4$,

	Sitting	Walking
Bezel Rotation	0.02 (SD = 0.08)	0.03 (SD = 0.11)
LL-Tapping	0.03 (SD = 0.09)	0.06 (SD = 0.17)
MS-Tapping	0.03 (SD = 0.11)	0.04 (SD = 0.13)
Swiping	0.07 (SD = 0.16)	0.09 (SD = 0.18)
V-Button	0.04 (SD = 0.11)	0.04 (SD = 0.11)

Table 5: Mean error values for the input techniques along the sitting and walking conditions with numeric input.
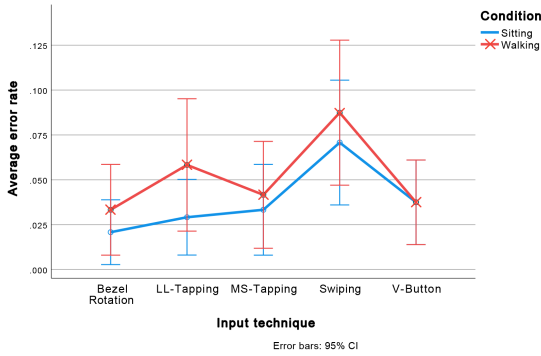
	Sitting	Walking
Bezel Rotation	0.03 (SD = 0.16)	0.05 (SD = 0.22)
LL-Tapping	0.04 (SD = 0.19)	0.09 (SD = 0.28)
MS-Tapping	0.01 (SD = 0.11)	0.06 (SD = 0.24)
Swiping	0.13 (SD = 0.33)	0.10 (SD = 0.30)
V-Button	0.03 (SD = 0.16)	0.04 (SD = 0.19)

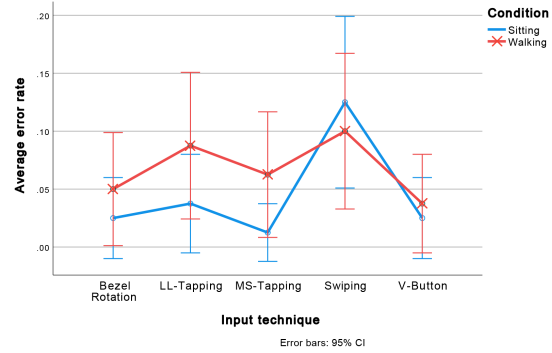**Table 6: Mean error rates for the input techniques for the sitting and walking conditions with text input.**

Figure 8: Error rates in the sitting and walking conditions per input techniques for numeric input.

Figure 9: Error rates in the sitting and walking conditions per input technique for list selection tasks

316) = 3.460, Partial $\eta2$ = 0.042). However post-hoc analysis with a Bonferroni correction revealed no significant difference in any of the pairwise comparisons (p > 0.05). Furthermore, the repeated measures ANOVA indicated a non-significant interaction effect (p = 0.857, F(4, 316) = 0.331), Partial $\eta2$ = 0.004), too. In other words, the differences in accuracy between the input techniques were independent of the activity level.

5.2.4 Effect of activity level and input technique on error rates for the list selection tasks. The mean error rate values for the activity level (i.e., sitting and walking) with text input are calculated as the number of false responses over the total number of responses, and the comparison between the five input techniques are shown in Table 6. We note how MS-Tapping resulted in the lowest error rate in list selection tasks, while Swiping resulted in the lowest error rates in the sitting condition. In the walking condition, V-Button had the lowest error rate and Swiping the highest. Fig. 9 displays the average error rates for the five input techniques. The separate lines indicate the level of activity (i.e., sitting and walking).

A two-way repeated measures ANOVA did not find a significant main effect of the activity level (p = 0.171, F(1, 79) = 1.905, Partial $\eta2$ = 0.024), suggesting that overall participants were equally precise whether seated or walking. The main effect of the input technique was significant (p = 0.009, F(4, 316) = 3.435, Partial $\eta2$ = 0.042). However, post-hoc analysis with Bonferroni correction found no significant difference in any of the pairwise comparisons (p > 0.05). Furthermore, the ANOVA indicated a non-significant interaction effect (p = 0.481, F(4, 316) = 0.872), Partial $\eta2$ = 0.011). In other

words, the differences in accuracy between the input techniques were independent of the activity level.

5.3 Qualitative data

Based on participant's feedback, LL-tapping emerged as the most favored method, with 31 participants expressing a preference for it. Conversely, Bezel rotation was the least popular, garnering only 8 participants' approval. More than half of the users found interaction methods that displayed one option at a time, such as Bezel rotation, Swipe, and Virtual button, to be less efficient.

For the bezel rotation, participants had diverse opinions. Among them, 21 participants indicated that during the sliding process, their fingers would obstruct the screen, while 17 participants found the need to slide their entire hand on the bezel to be cumbersome. Additionally, 22 participants mentioned their unfamiliarity with this interaction method, finding it challenging to use and often sliding past the intended answer. However, among the 6 participants who had prior experience with Samsung smartwatches, 5 of them displayed greater acceptance of this method. Furthermore, 8 participants considered it innovative and appealing. Overall, Bezel rotation received the least positive response, with only 8 participants expressing a liking for it.

Regarding MS-tapping participants' opinions varied. Fourteen participants believed that proper categorization could lead to faster identification of answer options, while 22 participants found the categorization process burdensome. Additionally, 9 participants mentioned the need for double tapping, requiring additional effort.

LL-tapping (single-choice categorization) received widespread appreciation, with 47 participants finding it efficient in quickly

displaying multiple options. However, 8 participants expressed concerns about accidentally scrolling past options and being unaware of missed answer choices. Thirty-seven participants expressed familiarity with this method and displayed greater acceptance.

Regarding the swiping method, 24 participants mentioned their familiarity with this interaction from mobile devices. Additionally, 9 participants felt that the smartwatch provided sufficient screen space for swiping between options. However, 42 participants experienced slow performance while swiping on the experimental smartwatch. Moreover, 45 participants considered the Swipe method, similar to the Virtual button, inefficient due to the display of only one option at a time.

The Virtual button method received a moderate level of acceptance, with 32 participants finding it fast, convenient, and associated with low error rates for switching between options. However, 45 participants expressed the need for continuous button presses to view all the options, considering it clumsy. Additionally, 5 participants found the buttons inconspicuous and too small, making them difficult to press.

6 DISCUSSION

Recognizing the potential of smartwatches to support ESM studies, this research has examined user interfaces for self-reporting that rely on touch-based input on smartwatches. We focused on two input tasks that are typical for ESM surveys: entering a number and list selection. We identified 13 broad categories of input techniques presented in related literature that can be distinguished in terms of the user interactions they involve. We implemented five of these to support the above two tasks and compared how users perform in terms of reaction times and error rates. Our comparison considered two different levels of activity, sitting and walking. Our first experiment compared three input techniques for the task of numeric input: Bezel Rotation, MS-tapping, and Swiping. The second experiment extended the comparison to include two more input techniques, namely LL-Tapping and V-button (see Fig. 5). All five techniques were adapted to support both tasks thus allowing the input of either numeric or categorical data without any inherent order, e.g., choosing among different states or activities. Next to looking for overall comparisons of different activity levels and different input techniques, the design of our experiments reflected our interest in uncovering interaction effects, which would indicate that the five input techniques are differently suited for different activity levels. Such a finding would also suggest a direction for future research, to develop adaptive user interfaces for wESM, where different input styles might be offered for the same task depending on the activity of the user.

Earlier research regarding the usability of input techniques for smartwatches has not attempted to assess the effects of different levels of activity on the accuracy and speed of input. Contrary to our expectations we found that user performance in terms of reaction times and error rates was similar when seated or walking. There may be different explanations for this finding. The most straightforward would be that users can equally well provide touch-based input on smartwatches while seated or walking, at least with the five techniques we assessed. Of course, different levels of activity which we did not include in our experiments, e.g., running, could

have a larger impact on usability but this does not appear relevant for ESM - researchers are likely to avoid surveying participants when they are engaged in intense physical activity or if necessary, opt for different input techniques that do not distract respondents or put them at risk. An alternative explanation could be that our measurements were not sensitive enough to these differences and larger samples or usage in different physical and social contexts could reveal differences in the performance of users.

In both studies and in different situations and for all comparisons we made, we found a significant main effect of the input technique. This confirms the intuition that not all user interfaces are as good for the two tasks studied. Regarding the reaction times our pilot study suggests the superiority of MS-tapping which is a multi-step input technique. Numbers are grouped into ranges and strings are grouped into categories, (which though supposes such groupings make sense logically), and can be accessed hierarchically: respondents first select the category and then an option within this category. In our study, we only examined such selection with two levels, as it is unusual for ESM studies to require choosing from extensive collections of items or ranges of numbers. This result is consistent with earlier findings for other interaction tasks, e.g., zoomable keyboards have been found effective for text input tasks [40, 41] and circular menus where a sub-range of potential responses is selected through a bezel rotation interface before an option within this sub-range is selected by tapping [31].

The Bezel rotation turned out to be the slowest of input techniques. The results of the second study were largely similar regarding the comparison among Bezel Rotation, Swiping, and MS-Tapping. However, the inclusion of the two other input techniques and extending our experimental tasks to include list selection led to further insights. Here too, the Bezel Rotation was found to be the slowest, but LL-Tapping (scrolling options by dragging the finger up or down and then tapping to select), was found to be the fastest way to enter numeric responses.

Regarding the accuracy of user input, we could not find a difference in error rates between sitting and walking. On the other hand, we noted how the Bezel Rotation which was the slowest also led to fewer errors both for text input and for list selection. LL-Tapping and MS-Tapping were the fastest but not as accurate. This could suggest that when designers are interested in reducing participant burden, which is an important concern in ESM studies, e.g., see [42], they could avoid the Bezel Rotation and choose one of the tapping interfaces. When the sampling protocol is sparser and the accuracy of responses more crucial for the research study, the Bezel Rotation input could be preferable.

Our expectations regarding interaction effects, which would suggest that the comparison favoring an input technique would be different when seated or when walking, were not confirmed, so unless future research uncovers interaction effects in different combinations of activity and input techniques, a guideline that can be offered to designers is to keep it simple and not pursue such an adaptation.

7 LIMITATIONS AND FUTURE WORK

Researchers have noted how circular and rectangular user interfaces on smartwatches may need to be approached differently to optimize

the use of the limited screen real estate [31]. Similar comparative evaluations also need to be done with rectangular watches, with user interfaces suited for that form factor. Our studies took place in a busy office space rather than in a well-controlled environment. While studies of input techniques on smartwatches have often been done in a more controlled environment, see for example [18, 28, 31], a real-life office environment can be more representative of the situations in which users may be prompted for a wESM study, with the obvious caveat that extraneous variables may have influenced our findings. Such extraneous variables may also include the relative familiarity with the hardware and the input techniques, which can be addressed by training participants before assessing their performance. Reducing environmental influences in a controlled lab setting may also help eliminate confounding factors. On the other hand, future field studies could extend the span of the experiment in time and space, allowing us to introduce realism and variation in the situations and moments chosen to prompt users. Taking this line of reasoning further, we note that our study does not capture the diversity of contexts in which users may find themselves when prompted to self-report in a wESM study. The pilot was conducted in an office space and the main study was in a library. Future research could consider different contexts such as outdoors, during transport, or at home. Also, they could be extended with more techniques for user input that we have not yet assessed in our study. For example, speech-based input is not an obvious choice for supporting wESM, as users are not always able or willing to speak when prompted in different situations. However, having the choice to use it in some situations may be helpful for enhancing response rates and the eventual user experience. Future studies could also include quantitative self-report measures of workload to rigorously compare the effort required from users for responding to wESM surveys with different input techniques. Furthermore, given that study participants were all young adults, future studies should aim to cover different age groups who may differ systematically in their performance and preferences.

8 CONCLUSION

We have examined touch-based input techniques to support the task of responding to ESM surveys on a smartwatch with numeric input and list selection. We implemented five input techniques that make use of the touchscreen of round smartwatches: MS-Tapping, Bezel Rotation, Swiping, Long List tapping, and Virtual Buttons for scrolling, which were compared with regard to the time it takes users to respond and the number of mistakes they make while doing so.

We found that MS-Tapping is the fastest technique, where users first select a segment within the range of options available for entering numbers or strings and then tap to select the preferred one within this range. The slowest, but most accurate technique turned out to be Bezel Rotation, where a virtual touch dial organized around the rim of the watch display allows one to flick between different options, and then tap in the central area to select that option. The activity level (i.e., sitting or walking) did not seem to affect the user's performance nor did it make much of a difference for which input technique works best, meaning that no adaptation is necessary with regards to the user activity. Rather, designers

should decide whether reducing participant burden or emphasizing the accuracy of the response of the user is their priority and select appropriate controls.

REFERENCES

[1] Rebecca Adaimi, Ka Tai Ho, and Edison Thomaz. 2020. Usability of a hands-free voice input interface for ecological momentary assessment. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 1–5.

[2] Anja Bachmann, Christoph Klebsattel, Andrea Schankin, Till Riedel, Michael Beigl, Markus Reichert, Philip Santangelo, and Ulrich Ebner-Priemer. 2015. Leveraging smartwatches for unobtrusive mobile ambulatory mood assessment. In *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers (UbiComp/ISWC'15 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 1057–1062. https://doi.org/10.1145/2800835.2800960

[3] Daniel J Barrett and DJ Barrett. 2005. ESP, the experience sampling program. *Retrieved March* 1 (2005), 2007.

[4] Chen Chen, Khalil Mrini, Kemeberly Charles, Ella Lifset, Michael Hogarth, Alison Moore, Nadir Weibel, and Emilia Farcas. 2021. Toward a Unified Metadata Schema for Ecological Momentary Assessment with Voice-First Virtual Assistants. In *Proceedings of the 3rd Conference on Conversational User Interfaces (CUI '21)*. Association for Computing Machinery, New York, NY, USA, 1–6. https://doi.org/10.1145/3469595.3469626

[5] Sunny Consolvo and Miriam Walker. 2003. Using the experience sampling method to evaluate ubicomp applications. *IEEE pervasive computing* 2, 2 (2003), 24–31.

[6] Anind K. Dey, Katarzyna Wac, Denzil Ferreira, Kevin Tassini, Jin-Hyuk Hong, and Julian Ramos. 2011. Getting closer: an empirical investigation of the proximity of user to their smart phones. In *Proceedings of the 13th international conference on Ubiquitous computing (UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 163–172. https://doi.org/10.1145/2030112.2030135

[7] Franz Faul, Edgar Erdfelder, Albert-Georg Lang, and Axel Buchner. 2007. G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods* 39, 2 (May 2007), 175–191. https://doi.org/10.3758/BF03193146

[8] Denzil Ferreira, Jorge Goncalves, Vassilis Kostakos, Louise Barkhuus, and Anind K Dey. 2014. Contextual experience sampling of mobile application micro-usage. In *Proceedings of the 16th international conference on Human-computer interaction with mobile devices & services*. 91–100.

[9] BJ Fogg. 2009. A behavior model for persuasive design. In *Proceedings of the 4th International Conference on Persuasive Technology (Persuasive '09)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/1541948.1541999

[10] Pegah Hafiz and Jakob Eyvind Bardram. 2020. The ubiquitous cognitive assessment tool for smartwatches: design, implementation, and evaluation study. *JMIR mHealth and uHealth* 8, 6 (2020), e17506.

[11] Javier Hernandez, Daniel McDuff, Christian Infante, Pattie Maes, Karen Quigley, and Rosalind Picard. 2016. Wearable ESM: differences in the experience sampling method across wearable devices. In *Wearable ESM: differences in the experience sampling method across wearable devices*. 195–205. https://doi.org/10.1145/2935334.2935340

[12] Stephen Intille, Caitlin Haynes, Dharam Maniar, Aditya Ponnada, and Justin Manjourides. 2016. microEMA: Microinteraction-based ecological momentary assessment (EMA) using a smartwatch. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '16)*. Association for Computing Machinery, New York, NY, USA, 1124–1128. https://doi.org/10.1145/2971648.2971717

[13] Vassilis-Javed Khan, Panos Markopoulos, Berry Eggen, Wijnand Ijsselsteijn, and Boris de Ruyter. 2008. Reconexp: a way to reduce the data loss of the experiencing sampling method. In *Proceedings of the 10th international conference on Human computer interaction with mobile devices and services*. 471–476.

[14] Alireza Khanshan, Pieter Van Gorp, and Panos Markopoulos. 2023. Experiencer: An Open-Source Context-Sensitive Wearable Experience Sampling Tool. In *Pervasive Computing Technologies for Healthcare*, Athanasios Tsanas and Andreas Triantafyllidis (Eds.). Springer Nature Switzerland, Cham, 315–331. https://doi.org/10.1007/978-3-031-34586-9_21

[15] Alireza Khanshan, Pieter Van Gorp, Raoul Nuijten, and Panos Markopoulos. 2021. Assessing the Influence of Physical Activity Upon the Experience Sampling

Response Rate on Wrist-Worn Devices. *International Journal of Environmental Research and Public Health* 18, 20 (2021). https://doi.org/10.3390/ijerph182010593

[16] Matin Kheirkhahan, Sanjay Nair, Anis Davoudi, Parisa Rashidi, Amal A. Wanigatunga, Duane B. Corbett, Tonatiuh Mendoza, Todd M. Manini, and Sanjay Ranka. 2019. A smartwatch-based framework for real-time and online assessment and mobility monitoring. *Journal of Biomedical Informatics* 89 (Jan. 2019), 29–40. https://doi.org/10.1016/j.jbi.2018.11.003

[17] Young-Ho Kim, Diana Chou, Bongshin Lee, Margaret Danilovich, Amanda Lazar, David E. Conroy, Hernisa Kacorri, and Eun Kyoung Choe. 2022. MyMove: Facilitating Older Adults to Collect In-Situ Activity Labels on a Smartwatch with Speech. , 21 pages. https://doi.org/10.1145/3491102.3517457

[18] Andreas Komninos and Mark Dunlop. 2014. Text input on a smart watch. *IEEE Pervasive Computing* 13, 4 (2014), 50–58.

[19] Yuki Kubo, Ryosuke Takada, Buntarou Shizuki, and Shin Takahashi. 2017. Exploring context-aware user interfaces for smartphone-smartwatch cross-device interaction. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–21.

[20] Sahiti Kunchay and Saeed Abdullah. 2020. WatchOver: using Apple watches to assess and predict substance co-use in young adults. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp-ISWC '20)*. Association for Computing Machinery, New York, NY, USA, 488–493. https://doi.org/10.1145/3410530.3414373

[21] Charlotte Rouzaud Laborde, Erta Cenko, Mamoun T. Mardini, Subhash Nerella, Matin Kheirkhahan, Sanjay Ranka, Roger B. Fillingim, Duane B. Corbett, Eric Weber, Parisa Rashidi, and Todd Manini. 2021. Satisfaction, Usability, and Compliance With the Use of Smartwatches for Ecological Momentary Assessment of Knee Osteoarthritis Symptoms in Older Adults: Usability Study. *JMIR Aging* 4, 3 (July 2021), e24553. https://doi.org/10.2196/24553 Company: JMIR Aging Distributor: JMIR Aging Institution: JMIR Aging Label: JMIR Aging Publisher: JMIR Publications Inc., Toronto, Canada.

[22] Reed Larson and Mihaly Csikszentmihalyi. 2014. The Experience Sampling Method. In *Flow and the Foundations of Positive Psychology: The Collected Works of Mihaly Csikszentmihalyi*, Mihaly Csikszentmihalyi (Ed.). Springer Netherlands, Dordrecht, 21–34. https://doi.org/10.1007/978-94-017-9088-8_2

[23] Luis A. Leiva, Alireza Sahami, Alejandro Catala, Niels Henze, and Albrecht Schmidt. 2015. Text Entry on Tiny QWERTY Soft Keyboards. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 669–678. https://doi.org/10.1145/2702123.2702388

[24] Todd Matthew Manini, Tonatiuh Mendoza, Manoj Battula, Anis Davoudi, Matin Kheirkhahan, Mary Ellen Young, Eric Weber, Roger Benton Fillingim, and Parisa Rashidi. 2019. Perception of Older Adults Toward Smartwatch Technology for Assessing Pain and Related Patient-Reported Outcomes: Pilot Study. *JMIR mHealth and uHealth* 7, 3 (March 2019), e10044. https://doi.org/10.2196/10044 Company: JMIR mHealth and uHealth Distributor: JMIR mHealth and uHealth Institution: JMIR mHealth and uHealth Label: JMIR mHealth and uHealth Publisher: JMIR Publications Inc., Toronto, Canada.

[25] Panos Markopoulos, Nikolaos Batalas, and Annick Timmermans. 2015. On the use of personalization to enhance compliance in experience sampling. In *Proceedings of the European Conference on Cognitive Ergonomics 2015*. 1–4.

[26] Stephen Oney, Chris Harrison, Amy Ogan, and Jason Wiese. 2013. ZoomBoard: a diminutive qwerty soft keyboard using iterative zooming for ultra-small devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, New York, NY, USA, 2799–2802. https://doi.org/10.1145/2470654.2481387

[27] Kyeongjin Park, Meuel Jeong, and Kyungdoh Kim. 2020. Usability evaluation of menu interfaces for smartwatches. *Journal of Computer Information Systems* 60, 2 (2020), 156–165.

[28] Sungkyu Park, Marios Constantinides, Luca Maria Aiello, Daniele Quercia, and Paul Van Gent. 2020. WellBeat: A Framework for Tracking Daily Well-Being Using Smartwatches. *IEEE Internet Computing* 24, 5 (Sept. 2020), 10–17. https://doi.org/10.1109/MIC.2020.3017867 Conference Name: IEEE Internet Computing.

[29] Veljko Pejovic, Neal Lathia, Cecilia Mascolo, and Mirco Musolesi. 2016. Mobile-based experience sampling for behaviour research. *Emotions and personality in personalized services: Models, evaluation and applications* (2016), 141–161.

[30] Simon T Perrault, Eric Lecolinet, James Eagan, and Yves Guiard. 2013. Watchit: simple gestures and eyes-free interaction for wristwatches and bracelets. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1451–1460.

[31] Katrin Plaumann, Michael Müller, and Enrico Rukzio. 2016. CircularSelection: optimizing list selection for smartwatches. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*. 128–135.

[32] Aditya Ponnada, Caitlin Haynes, Dharam Maniar, Justin Manjourides, and Stephen Intille. 2017. Microinteraction Ecological Momentary Assessment Response Rates: Effect of Microinteractions or the Smartwatch? *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (Sept. 2017), 92:1–92:16. https://doi.org/10.1145/3130957

[33] Gulnar Rakhmetulla and Ahmed Sabbir Arif. 2022. SwipeRing: Gesture Typing on Smartwatches Using a Segmented Qwerty Around the Bezel. https://openreview.net/forum?id=eVyaNs3sm8a

[34] Samsung. 2023. About | Tizen. https://www.tizen.org/about

[35] Samsung. 2023. Galaxy Watch Active2 Smartwatch. https://www.samsung.com/nl/wearables/galaxy-watch-active2-under-armour-40mm/SM-R830NZKUXEN/

[36] Stefan Schneider, Doerte U Junghaenel, Joshua M Smyth, Cheng K Fred Wen, and Arthur A Stone. 2023. Just-in-time adaptive ecological momentary assessment (JITA-EMA). *Behavior Research Methods* (2023), 1–19.

[37] Sojeong Seong, Soyeon Park, Yong Han Ahn, and Heejung Kim. 2022. Development of an integrated fatigue measurement system for construction workers: a feasibility study. *BMC Public Health* 22, 1 (Aug. 2022), 1593. https://doi.org/10.1186/s12889-022-13973-5

[38] Saul Shiffman, Arthur A. Stone, and Michael R. Hufford. 2008. Ecological Momentary Assessment. *Annual Review of Clinical Psychology* 4, 1 (2008), 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415 _eprint: https://doi.org/10.1146/annurev.clinpsy.3.022806.091415.

[39] Gaganpreet Singh, William Delamare, and Pourang Irani. 2018. D-SWIME: A Design Space for Smartwatch Interaction Techniques Supporting Mobility and Encumbrance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174208

[40] Keiichi Ueno, Kentaro Go, and Yuichiro Kinoshita. 2016. Design and Evaluation of EdgeWrite Alphabets for Round Face Smartwatches. In *Adjunct Proceedings of the 29th Annual ACM Symposium on User Interface Software and Technology (UIST '16 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 177–178. https://doi.org/10.1145/2984751.2984757

[41] Keiichi Ueno, Kentaro Go, and Yuichiro Kinoshita. 2017. Investigation of smartwatch touch behavior with different postures. In *Proceedings of the 16th International Conference on Mobile and Ubiquitous Multimedia (MUM '17)*. Association for Computing Machinery, New York, NY, USA, 57–61. https://doi.org/10.1145/3152832.3152843

[42] Niels van Berkel, Denzil Ferreira, and Vassilis Kostakos. 2017. The Experience Sampling Method on Mobile Devices. *Comput. Surveys* 50, 6 (Dec. 2017), 93:1–93:40. https://doi.org/10.1145/3123988

[43] Selina Volsa, Bernad Batinic, and Stefan Stieger. 2022. Self-Reports in the Field Using Smartwatches: An Open-Source Firmware Solution. *Sensors* 22, 5 (Jan. 2022), 1980. https://doi.org/10.3390/s22051980 Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[44] Cheng Xu and Kent Lyons. 2015. Shimmering smartwatches: Exploring the smartwatch design space. In *Proceedings of the Ninth international conference on tangible, embedded, and embodied interaction*. 69–76.

[45] Xinghui Yan, Shriti Raj, Bingjian Huang, Sun Young Park, and Mark W. Newman. 2020. Toward Lightweight In-situ Self-reporting: An Exploratory Study of Alternative Smartwatch Interface Designs in Context. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (Dec. 2020), 158:1–158:22. https://doi.org/10.1145/3432212

[46] Hui-Shyong Yeo, Juyoung Lee, Andrea Bianchi, and Aaron Quigley. 2016. WatchMI: applications of watch movement input on unmodified smartwatches. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct (MobileHCI '16)*. Association for Computing Machinery, New York, NY, USA, 594–598. https://doi.org/10.1145/2957265.2961825

A   NUMERIC AND TEXT INPUT QUESTIONS IN THE MAIN STUDY

1	Can you insert the number 5?
2	How many fingers do you have on your right hand?
3	2021 is how many years ago?
4	2 + 6 = ?
5	3 - 2 = ?
6	1 + 6 = ?
7	How many toes do you have on your left foot?
8	Can you insert the number 7?
9	Can you insert the number 10?
10	Can you insert the number 8?
11	Can you insert the number 3?
12	3 + 6 = ?
13	1 + 1 = ?
14	How many thumbs do you have?
15	Can you insert number 2?

Table 7: Items used for the numeric input questions.

1	Which country is famous for the Pyramids?
2	What country did Van Gogh come from?
3	Which country is known for the Eiffel Tower?
4	What is the country where the Great Wall is located?
5	Tokyo is the capital of which country?

Table 8: Items used for text input questions.